VOLUME 23 · NUMBER 24 · AUGUST 20 2005

JOURNAL OF CLINICAL ONCOLOGY

SPECIAL ARTICLE

Preliminary Data Release for Randomized Clinical Trials of Noninferiority: A New Proposal

Edward L. Korn, Sally Hunsberger, Boris Freidlin, Malcolm A. Smith, and Jeffrey S. Abrams

From the Biometric Research Branch and Clinical Investigations Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD.

Submitted October 7, 2004; accepted February 2, 2005.

Authors' disclosures of potential conflicts of interest are found at the end of this article.

Address reprint requests to Edward L. Korn, PhD, Biometric Research Branch, EPN-8128, National Cancer Institute, Bethesda, MD 20892; e-mail: korne@ctep.nci.nih.gov.

0732-183X/05/2324-5831/$20.00

DOI: 10.1200/JCO.2005.02.105

ABSTRACT

Noninferiority trials often require a long follow-up period for the data to reach the maturity needed for definitive analysis. A proposal is presented that allows for early release of outcome data from a carefully specified subset of noninferiority trials. This subset is defined so that the early release of the data will be potentially useful to patients who face a treatment decision but will not compromise the integrity of the trial or interfere with the completion of the trial to its definitive analysis. In particular, the release of the data will only occur after the last participant has been randomly assigned and is off treatment-arm–specific therapy and only if it is unlikely that subsequent treatment and/or follow-up practices will change based on the knowledge of released data. In contrast to standard interim monitoring, (1) the release of the data would be automatic and independent of the observed data, and (2) the trial would continue on to its planned final analysis and not be stopped. Examples are given demonstrating how the proposal would work, along with a discussion of possible objections to the proposal.

J Clin Oncol 23:5831-5836.

INTRODUCTION

Randomized clinical trials are the gold standard for obtaining unbiased comparisons of different treatment regimens. Unfortunately, they sometimes take a long time to complete, which tends to be a problem when the primary outcome is a time-to-event outcome (such as overall survival or time to recurrence) and the event rate is low. It can be an especially serious problem for noninferiority trials, in which the targeted difference between the treatment arms (representing inferiority) may be small, leading to a large number of events being required for the trial to obtain a reliable conclusion. This often results in a prolonged follow-up period to allow for evidence to mature to the level necessary for definitive analysis. Noninferiority trials are used frequently to compare a standard treatment with a modification of the standard treatment that is expected to have less toxicity and/or morbidity (eg, by reducing chemotherapy or radiotherapy or by using a less extensive surgery). Frequently in these situations the treatments being compared in the trial are already being used in the community. Depending on the results of the trial, waiting years for definitive results from a trial means that many patients will be treated with an inferior therapy or, alternatively, a therapy that is more aggressive than necessary. During this waiting period, the data from the noninferiority study may represent the best available information on the relative efficacy of the treatments. In most instances, the only other data available are from noncontrolled, phase II or pilot clinical trials. We propose that, in specific circumstances, the risks of releasing preliminary outcome data from noninferiority trials are sufficiently small that consideration should be given to releasing these data widely so that they can be used by patients and physicians for treatment decisions.

Randomized trials have formal data-monitoring guidelines for release of outcome data that are applied by independent

data-monitoring committees. These guidelines ensure that if during the course of the trial the accruing results become definitive, then the trial will be stopped and the results will be disseminated. Early-stopping guidelines help with the problem of long trials, but because the guidelines are designed to preserve type I and II error rates (defined in the Appendix), the interim results of the trial must be extreme for a data-monitoring committee to stop a trial. Therefore, standard stopping guidelines only infrequently shorten the length of a trial and provide the public with early access to outcome data. In this article we suggest a strategy that may sometimes be used to provide early public access to reasonably reliable outcome data that may be helpful in making treatment decisions. We suggest that, in specific settings, early release of outcome data can be done without harm to the future conduct of the trial and without being misleading too often. In these settings, release of such data before the final planned study analysis can be in the public interest even when the data are not definitive. An important point to emphasize is that unlike standard monitoring guidelines for stopping trials, the proposed approach for release of the preliminary results is not data driven and it does not affect the error structure of the trial, because any study-level conclusions will be made on the basis of final results of the trial.

The details of the proposal and the types of trials in which early release can be justifiably proposed are described in the next section along with two examples. We end with a discussion of possible extensions and objections to the proposal. The Appendix contains a brief review of the design of noninferiority trials along with some technical justifications for the proposal.

PROPOSAL

For a noninferiority trial, the proposal is to release the preliminary data on a regular basis (eg, based on the number of events observed) starting after a fixed amount of time after the last patient has completed his or her randomly assigned treatment. Table 1 lists some conditions for applying this proposal. The first three conditions are necessary to ensure that release of the preliminary results will not harm the future conduct of the trial. If all patients have completed their randomly assigned therapy and if it is highly unlikely that either changes in subsequent treatment or follow-up intensity will result from public dissemination of preliminary results, then the research community can be reasonably confident that the integrity of the trial will be protected. In addition, note that the patients enrolled on the trial will not be affected by the early release of the trial data if these first three conditions are met.

Conditions four and five are included because early release of data is not innocuous, and therefore the benefits of this approach should outweigh any disadvantage. If there is not a relatively long follow-up period (condition four), then the time gained in releasing the data early may not counter sufficiently the possible confusion resulting from more than one set of results from the trial. Condition five states that both treatments should be available and used in the community. If the reduced therapy arm is available in the community, the primary risks are that it may be inferior (in terms of efficacy) to the standard therapy and that it may be adopted prematurely by the community. Access to reasonably reliable preliminary data may stem the adoption of the inferior therapy pending results from the final study analysis. On the other hand, if the community is only using the standard therapy, there is less impetus to report preliminary trial data, because community use of a standard therapy that is more aggressive than it needs to be is generally less of a problem than community use of a less morbid but inferior therapy.

Condition six in Table 1 is related to the fact that trial results for some special types of trials can be misleading if there is insufficient follow-up, even with a large number of events. For example, in a trial of aggressive surgery versus less aggressive surgery, the investigators might expect that the aggressive surgery arm initially will look inferior (because of surgical mortality), but its long-term survival will

Table 1. Conditions for Early Release of Data Before a Trial Is Stopped

1.	All patients have already been randomly assigned and are off treatment-arm–specific therapy
2.	It is very unlikely or impossible that patients would modify their subsequent treatment (if any) based on the knowledge of preliminary trial results
3.	It is very unlikely or impossible that the intensity of the follow-up for the primary outcome would be modified based on the knowledge of preliminary trial results
4.	The time from when the last patient is off his or her randomly assigned treatment to the time of the definitive analysis is relatively long
5.	The therapies on both treatment arms are available and being used in the community
6.	The planned analysis strategy of the trial is appropriate with reduced follow-up
7.	When preliminary results are reported, they will be clearly designated as being preliminary results with mention of when the final definitive results are expected to be reported
8.	Reported preliminary results should prominently display the uncertainty associated with any estimates of relative treatment efficacy (eg, with confidence intervals and the minimum and median follow-up time)
9.	In addition to providing the relative treatment efficacy in terms of the primary outcome, reporting of preliminary results should include all results from the trial relevant to decision making about the treatment (eg, toxicity profiles)

be better than the less aggressive surgery. In this situation of "crossing hazards," the usual log-rank test based on proportional hazards may not be appropriate, and this would be noted in the statistical analysis plan for the trial. Because sufficient follow-up time is required for meaningful results for this type of trial, we would not recommend releasing data early.

The final three conditions for early release relate to the manner in which these data are released. To minimize public confusion that may arise from the early release of outcome data, it is important that these data be clearly designated as "preliminary results" that are subject to change and that mention be made of when the final definitive results are expected to be reported, as well as when any other planned releases of preliminary results are expected. Public understanding may also be enhanced by prominently displaying the uncertainty associated with the estimates of relative treatment efficacy. For example, CIs for estimates of relative treatment efficacy and the minimum and median follow-up time should be provided, and efforts should be made to explain them in simple, lay terms.

With early release, any trial results that would be relevant to the treatment-making decision should be released along with the efficacy data (eg, toxicities of the various treatments). There should be acknowledgment, when appropriate, that important toxicities caused by some treatments may occur as late effects and hence would be noted only after longer follow-up. The usual descriptive data that would be presented in an analysis of definitive results (eg, proportions of patients receiving their randomly assigned treatment) should also be given in the release of the preliminary results. In addition, to avoid any erosion of the medical community's understanding of the confidential interim-analysis concept, the presentation of the results should state that the analysis and its public presentation were specified in the protocol. Although careful presentation is required, the preliminary results should be made widely available to the public through publications in journals and presentations at professional meetings.

We now address how long to wait before starting the release of preliminary data. The preliminary data could be misleading in two ways: they could suggest that (1) the treatments (standard S and reduced R) are equivalent when in fact R is inferior to S, or (2) R is inferior to S when in fact they are equivalent. Although early release of preliminary data may be helpful to the public, the data will be more immature and therefore potentially more misleading. We would like to choose a time to start releasing the data so that the probability of the data being misleading is acceptable even when they are not definitive. We recommend this time to be when one has observed 35% of the expected number of events at the end of the trial or when all patients have completed randomly assigned treatment, whichever comes later. The idea is to have a time point at which data will have reasonable reliability even when they are not definitive; a technical justification for this time point is given in the Appendix. We now present two examples to demonstrate the proposal.

Example One

Suppose the standard surgical procedure results in a 65% 5-year survival. A new, less extensive surgical procedure is beginning to be used in the community and seems to have survival similar to the standard procedure. However, because the experience is limited and based on historical comparisons, one cannot rule out a clinically important decrement in survival compared with the standard procedure. A randomized clinical trial is designed to answer definitively whether the new surgical procedure (treatment R) is inferior to the standard surgical procedure (treatment S). It is decided that if the true hazard ratio of R to S is ≤ 1.15, then R should be considered not inferior to S. Using an exponential assumption, this corresponds to a 5-year survival of 60.9% for arm R. It is estimated that 700 patients per year would accrue to this trial.

We desire our trial to have the following operating characteristics: if the true hazard ratio is ≥ 1.15, then the probability of declaring noninferiority is $\leq 5\%$; if the treatments are perfectly equivalent (true hazard ratio $= 1$), then the probability of declaring R inferior to S is $\leq 10\%$. To accomplish this goal, 2,800 patients are accrued over 4 years, and the final analysis is planned for 10 years after that. Interim-analysis guidelines for stopping the trial are discussed in the Appendix.

To apply our proposal we note that at the final analysis at the end of the trial, an expected 1,850 events (deaths) will have occurred. Therefore, we would release the preliminary data when 648 (35%) of 1,850 events have been observed. This would be expected to happen approximately 1 year after the last patient has been randomly assigned. With such a long follow-up time, it is reasonable to have two more early releases of data. Therefore, we would also release the data when 60% and 80% of the events have been observed, corresponding to approximately 4 and 6.5 years after the last patient has been randomly assigned.

Example Two

Suppose a randomized clinical trial is designed to demonstrate whether a reduced radiation therapy (treatment R) is as good as the standard radiation therapy (treatment S) for a certain subset of node-negative breast cancer patients. Assume that the 5-year disease-free survival for the standard arm is expected to be 90%. A clinically meaningful decrement in disease-free survival is taken to be a 5-year disease-free survival of 87.2%, which corresponds to a hazard ratio of 1.3 (using an exponential assumption). Three thousand two hundred patients will be accrued over 5 years. The final analysis will take place when there are 530 events, which should be approximately 5 years after accrual is done.

The design has the following properties: If the true hazard ratio is ≥ 1.3, then the probability of declaring noninferiority is $\leq 5\%$; if the treatments are perfectly equivalent, then the probability of declaring R inferior to S is $\leq 10\%$. Interim-analysis guidelines for stopping the trial are discussed in the Appendix.

To apply our proposal, we would release the data when there were 186 (35%) of 530 events, or when the last patient is off randomly assigned treatment, whichever comes later, and when there were 366 (69%) of 530 events. This should occur approximately 5 and 7.5 years after commencement of the trial, respectively.

DISCUSSION

For noninferiority trials, which may have follow-up periods extending as long as 10 years, there is strong impetus for developing mechanisms for providing patients and physicians with access to outcome data before the final study analysis. However, it is essential that this be done in ways that protect the integrity of the study and in ways that maximize public understanding of the preliminary nature of the results. We suggest that application of the eight conditions outlined in Table 1 may be sufficient for assuring study protection and enhancing public understanding.

A possible extension of our proposal would be to apply it to one-sided superiority trials. For example, the new therapy might consist of a standard therapy plus the addition of another agent. The conditions listed in Table 1 would still need to apply. There likely will be only a small number of trials for which these eight conditions are met. For example, the proposal would not be applicable to trials involving new drugs, because the therapies involving the new drugs would not be available or used in the community (condition five). Even for a trial involving a drug therapy that is being used in the community, the proposal would not be appropriate if there was the possibility that patients enrolled on the study would take the drug therapy if preliminary results suggested that it was efficacious (condition two). The proposal would be attractive for some adjuvant trials, because the follow-up can be long, but in many of these trials, patients will be on treatment-arm–specific therapy for a long time, making the proposal less applicable (condition one). A different possible extension to our proposal would be to apply it to a nonrandomized trial designed to test whether a treatment is noninferior to a control treatment for which historical data are available. Such nonrandomized trials are not uncommon in pediatric oncology.

We end by noting some arguments against the proposal. One argument is that if preliminary data offer a reliable conclusion that patients and physicians can use in making treatment decisions, then either the interim monitoring guidelines for stopping the trial will apply or the trial was designed to be unnecessarily large or long. If the preliminary data do not offer a reliable conclusion, then releasing them will not be helpful. We are sympathetic to this argument but note that there are different degrees of reliability. For example, suppose the results from example two were released after 35% of the total number of events were observed. Consider a patient with node-negative breast cancer faced with a treatment decision. Although specific conditions may vary from patient to patient, one of the questions for the patient is whether the reduced morbidity of the reduced radiation therapy is offset by the potential increase in disease-free survival. Suppose the observed hazard ratio from the released data is approximately 1.0. One can calculate that a (one-sided) upper 90% confidence bound for the hazard ratio will be approximately 1.21 and the upper 80% confidence bound will be approximately 1.13. (No adjustment for multiple comparisons is required for these confidence bounds; see the Appendix.) Access to these data would reassure the patient that (1) according to the best available estimate (at that time) reduced radiation therapy has the same disease-free survival as the standard treatment, and (2) with reasonable certainty the decrease in 5-year disease-free survival would be from 90% to no less than 88.0% to 88.8%. On the other hand, suppose the observed hazard ratio from the early release of data was 1.3. The lower 90% (80%) confidence bound for the hazard would be approximately 1.08 (1.15). Access to these data would caution the patient that (1) according to the best available estimate there is a 30% increase in event rate with the reduced radiation therapy, and (2) with reasonable certainty disease-free survival is higher with the reduced radiation therapy. Of course, the observed hazard ratio may fall into an intermediate range in which case the data will be less useful for making treatment decisions. In all cases, we would suggest that reports of preliminary results include results from other relevant trials, although these results may be based on limited data.

Some other arguments against the proposal are related to the fact that it leads to multiple reporting of the results of a trial. Multiple reporting of results increases the workload at the statistics and data center responsible for the study, as typically clinical trials data will be "cleaned up" to assure that they are accurate and up-to-date before public release. Although resolution of eligibility issues and delivery-of-treatment data could be moved up to before the first release of preliminary data (from before the definitive analysis), updating of the outcome data will need to be done before each release of preliminary data and therefore require additional resources. Whatever additional resources are required should be provided by the organization funding the trial. Multiple releases of the data from a trial will also bring new challenges in presenting the information to health care providers and the public in a comprehensible manner.

Although it may be possible to use abstracts at the American Society of Clinical Oncology or other national/international disease-specific meetings as venues for publication of preliminary results, this may not address the difficulty that clinicians and patients will have in incorporating trial results that are changing frequently into their decision making. All these arguments related to multiple reporting can be somewhat mitigated if the preliminary data-reporting interval is lengthened. The trade-off is between the concerns expressed above and having the most up-to-date data available. This should be considered on a trial-by-trial basis, with the preliminary data-reporting interval written into the protocol for the trial (along with the standard interim monitoring for extreme results).

Finally, the proposal relies on the assumption that the trial will be completed successfully and the conclusions at the end of the trial will become the ones used for future treatment decisions. We recommend that this assumption be empirically tested with some trials before the proposal becomes widely adopted.

APPENDIX

Design and Analysis of Noninferiority Trials

There are two approaches to the design and analysis of noninferiority trials. One[1] uses a hypothesis-testing framework but switches the roles of the null and alternative hypotheses used in superiority trials, and the other uses a confidence-interval approach.[2,3] Both approaches yield approximately the same designs. Let Δ_c be the hazard ratio representing the smallest hazard ratio for which we would consider the treatment to be clinically inferior (eg, $\Delta_c = 1.3$). (There has been much discussion in the literature[4] about the choice of Δ_c to ensure that if treatment *R* is deemed noninferior to treatment *S*, then treatment *R* is better than no treatment. For the purposes of this article, we assume that an appropriate Δ_c has been chosen.) When the trial is over, there are two possible errors that can be made. A type I error occurs when we conclude that treatment *R* is not inferior to treatment *S* when in fact the hazard ratio is $\geq \Delta_c$. A type II error occurs when we conclude the treatment *R* is inferior to treatment *S* when in fact they are equivalent or *S* is actually inferior to *R*, that is, the hazard ratio is ≤ 1.0. (Note that the definitions of type I and II errors are switched from the definitions for superiority trials, because the null and alternative hypotheses have been switched.) Because a type I error is typically considered more serious than a type II error, designs usually require a smaller probability of a type I error than a type II error.

Because a conclusion will be stated concerning noninferiority if the trial stops early because of interim monitoring, interim monitoring for stopping must be taken into account when designing the trial and calculating the probabilities of type I and II errors. To keep the probabilities of type I and II errors below desired levels, various approaches have been developed for interim monitoring that stop the trial early for extreme events. We demonstrate two of these approaches in the context of the two examples given in the main article. We note that our proposal for release of preliminary data should not be viewed as an excuse for designing trials with excessively long follow-up, because the definitive analysis will be at the end of the follow-up (unless the trial stops early as a result of interim monitoring).

Example one (continued). A formal interim analysis is planned yearly, starting at year 5 of the trial (1 year after the end of accrual) when approximately 34% of the expected total of 1,850 events have been observed. At any interim analysis, if the 99% lower confidence bound for the hazard ratio Δ is > 1.0, the trial is stopped and treatment *R* is declared inferior to *S*. At each interim analysis the 99.5% upper confidence bound for Δ is also calculated. At any interim analysis, if this upper bound is < 1.15, the trial is stopped and *R* is declared to be not inferior to *S*. If the trial is not stopped early, the final analysis is performed at year 14: If the 91% lower confidence bound for Δ is > 1, then *R* is declared to be inferior to *S*; otherwise, *R* is declared to be noninferior to *S*. (If there had been no interim monitoring, one could use a 90% confidence bound at the final analysis and a slightly smaller sample size.) This design approximately has the probability of a type I error being $\leq 5\%$ and type II error being $\leq 10\%$. The unchanging confidence bounds at the interim analyses is similar to a Haybittle–Peto approach,[5,6] which is used for monitoring of superiority trials. With this approach, there is little statistical cost in formal monitoring quite often once it begins.[7]

Example two (continued). Three formal interim analyses are planned when there have been 133, 265, and 398 events observed. These analysis times correspond to 25%, 50%, and 75% of the total of 530 events that will be observed at the time of the final analysis. At the first interim analysis, if the one-sided log-rank test with a significance level of .005 rejects the hypothesis that the treatments are equivalent in favor of the hypothesis that *S* is better than *R*, then the trial is stopped, with treatment *R* declared inferior to *S*. For the second and third interim analyses, the same rule applies except the significance level of the tests are .019 and .051, respectively. If the trial is not stopped at any of these interim analyses, the final analysis should be approximately 5 years after the accrual is over. At the final analysis, if the one-sided log-rank test with a significance level of .082 rejects the hypothesis that the treatments are equivalent in favor of the hypothesis that *S* is better than *R*, then treatment *R* is declared inferior to *S*; otherwise, treatment *R* is declared noninferior to treatment *S*. This design approximately has the probability of a type I error being $\leq 5\%$ and type II error being $\leq 10\%$. These significance levels at the interim analyses are based on the truncated O'Brien–Fleming

approach,[8,9] which is used for monitoring superiority trials. In practice, the formal monitoring times may not be exactly when 133, 265, and 398 events have been observed, so a Lan–Demets spending function approach[10] would be used to determine the precise significance levels.

Justification of Timing of Data Release

A heuristic justification for beginning to release preliminary data when 35% of the total expected number of events at the end of the trial have been observed is as follows: At the end of the trial, there will be an observed hazard ratio calculated, $\hat{\Delta}$, and R will be declared noninferior to S if $\hat{\Delta} \leq \Delta_E$, where Δ_E depends on the trial-design parameter Δ_c. (This assumes the confidence-interval approach to analysis and that there is no formal monitoring for stopping.) Assuming the trial is designed with a type I error of 5% and type II error of 10%, we can calculate the probability that $\hat{\Delta} \leq \Delta_E$ when $\hat{\Delta}$ is based on an analysis done with 35% of the total expected number of events and when the true hazard ratio is Δ_c. This probability is approximately15%. That is, when R is inferior to S (by at least Δ_c), the probability of observing data at 35% events that suggests noninferiority ($\hat{\Delta} \leq \Delta_E$) is approximately $\leq$ 15%. If the true hazard ratio were substantially larger than the design parameter (eg, $1.5 \times \Delta_c$), then the probability of observing $\hat{\Delta} \leq \Delta_E$ could be substantially smaller (eg, 3.3%). Other choices of when to start releasing data early could be justified by other considerations. Note that the eventual presentation of the final results have the potential of correcting any misleading impressions left by the release of preliminary data. Therefore, one could argue that presentation of preliminary data that turn out to be misleading, although unfortunate, is not as significant of a problem as the presentation of final results that are misleading.

Multiple-Comparisons Issues and the Early Release of Data

As noted above, multiple looks at the data for interim monitoring for stopping a trial for extreme results need to be adjusted for in the statistical analysis. However, this is not true for the early reporting of data with our proposal. In particular, CIs for the results can be reported without a multiple-comparisons adjustment. There are three reasons for this. First, unlike early stopping, the results at the end of the trial are going to be reported, and that will be the definitive report. Second, unlike early stopping, the results will be released regardless of the size of the observed treatment effect at the time (ie, the release is not data driven). Finally, the primary audience for the early-released data are patients facing a treatment decision. These patients will use the most recently reported results rather than the "best" results of some set of analyses. For these reasons, no multiple-comparisons adjustment is required.

Authors' Disclosures of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

REFERENCES

1. Blackwelder WC: "Proving the null hypothesis" in clinical trials. Controlled Clin Trials 3:345-353, 1982
2. Jennison C, Turnbull BW: Interim analyses: The repeated confidence interval approach (with discussion). J R Stat Soc B 51:305-361, 1989
3. Durrleman S, Simon R: Planning and monitoring of equivalence studies. Biometrics 46: 329-336, 1990
4. Temple R, Ellenberg SS: Placebo-controlled trials and active-control trials in the evaluation of new treatments. Ann Intern Med 133:455-463, 2000
5. Haybittle JL: Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 44:793-797, 1971
6. Peto R, Pike MC, Armitage P, et al: Design and analysis of randomized clinical trials requiring prolonged observations of each patient: Introduction and design. Br J Cancer 34:585-612, 1976
7. Freidlin B, Korn EL, George SL: Data monitoring committees and interim monitoring guidelines. Control Clin Trials 20:395-407, 1999
8. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. Biometrics 35:549-556, 1979
9. Fleming TR, Harrington DP, O'Brien PC: Designs for group sequential tests. Control Clin Trials 5:348-361, 1984
10. Lan KKG, DeMets DL: Discrete sequential boundaries for clinical trials. Biometrika 70:659-663, 1983